Bioinformatics

Week 5， 2020-04-01

Xiaofan Zhou

BLAST

- **B**asic **L**ocal **A**lignment **S**earching **T**ool
  - *"(it) finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance"*.

Why BLAST?

- To find homologs

- To find regions that are conserved (or not)

- To infer structure or function

- ……

BLAST

- **Basic Local Alignment Searching Tool**
  - *"(it) finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance".*

- Pairwise sequence alignment
  - *"the process of lining up two sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences)".*
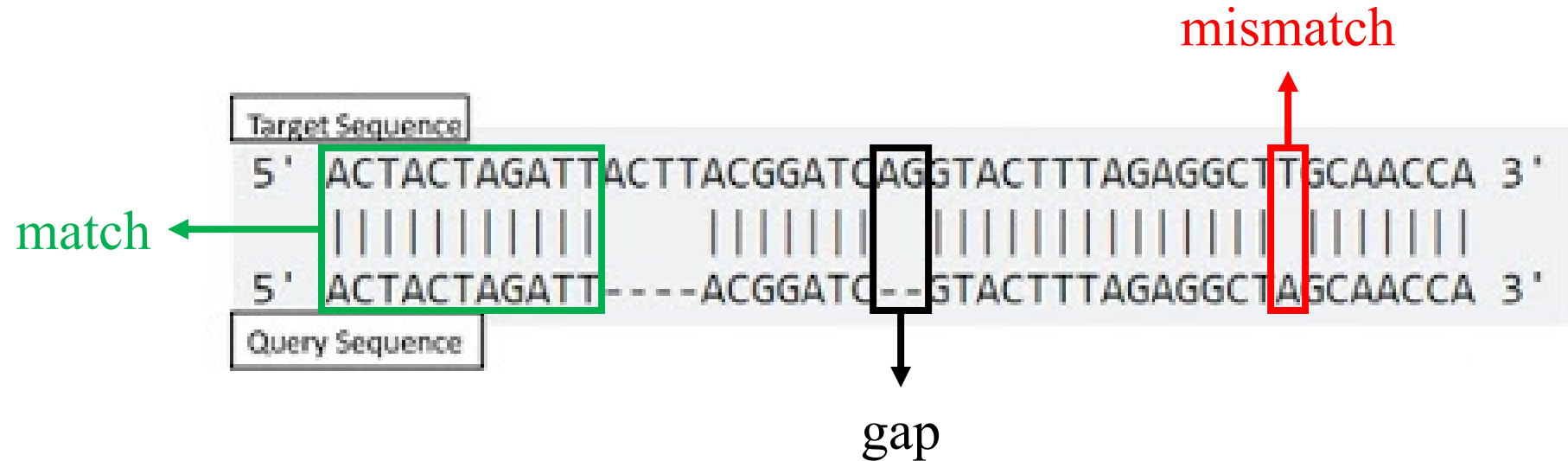
```
Target Sequence:
ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA

Query Sequence:
ACTACTAGATTACGGATCGTACTTTAGAGGCTAGCAACCA
```

Scoring of pairwise alignment: DNA

Default Mega-BLAST scoring matrix:
- match: +1
- mismatch: -2
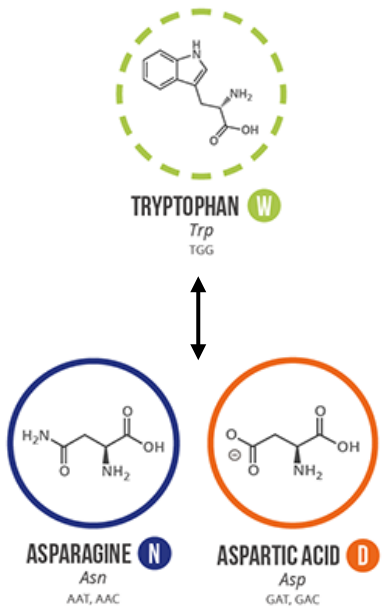- gap opening: 0
- gap extension: -2.5

The above alignment has 39 matches,
1 mismatch, 6 gaps (2 openings):
$$= 39*1 + 1*(-2) + 6*(-2.5) + 2*0$$
$$= 22$$

Protein scoring matrix

Protein scoring matrix

BLOSUM62, the default scoring matrix of BLASTP
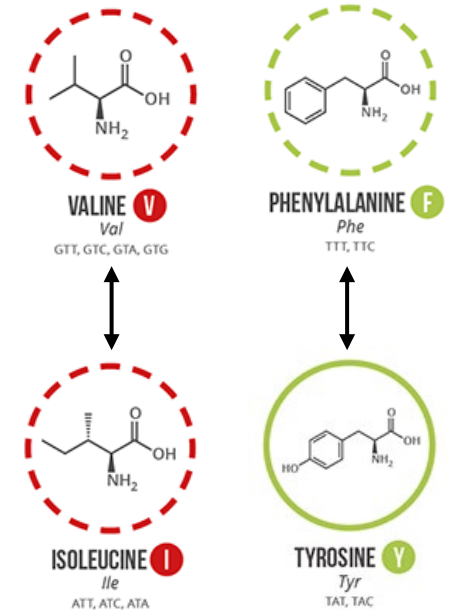
Scoring of pairwise alignment: protein

```
Score = 18.1 bits (35),  Expect = 0.015, Method: Composition-based stats.
Identities = 11/24 (45%), Positives = 12/24 (50%), Gaps = 2/24 (8%)

Query  12   VTALWGKVNVD--EVGGEALGRLL   33
            V  +WGKV  D     G E L RL
Sbjct  11   VLNVWGKVEADIPGHGQEVLIRLF   34


match        4  11 5   6     6 5 4 5      sum of matches: +60
                   6 4                4
mismatch    -1 1      0   -2 -2  -4  0    sum of mismatches: -13
               -2        0    -3   0
gap open              -11                 sum of gap penalties: -13
gap extend            -2

                                          total raw score: 61 - 13 - 13 = 35
```

Global vs. local alignment

global alignment – 15% identity

local alignment – 30% identity

Global alignment: Needleman-Wunsch

		G	C	C	C	T	A	G	C	G
	0	-2	-4							
G	-2	?								
C	-4									
G										
C										
A										
A										
T										
G										

G C
_ _

G _
C _

- match: +1
- mismatch: -1
- gap opening: 0
- gap extension: -2

Global alignment: Needleman-Wunsch

$\begin{array}{c} G \\ G \end{array}$ ➡ 1 $\begin{array}{c} G \\ - \end{array}$ ➡ $\begin{array}{c} G- \\ -G \end{array}$ ➡ -4

		G	C	C	C	T	A	G	C	G
		0	-2	-4						
G	-2	?							
C	-4								
G									
C									
A									
A									
T									
G									

$\begin{array}{c} -G \\ G- \end{array}$ -4

Global alignment: Needleman-Wunsch

		G	C	C	C	T	A	G	C	G
	0	-2	-4							
G	-2	1	-1							
C	-4	-1	2							
G										
C										
A										
A										
T										
G										

Global alignment: Needleman-Wunsch

		G	C	C	C	T	A	G	C	G
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
G	-2	1	-1	-3	-5	-7	-9	-11	-13	-15
C	-4	-1	2	0	-2	-4	-6	-8	-10	-12
G	-6	-3	0	1	-1	-3	-5	-5	-7	-9
C	-8	-5	-2	1	2	0	-2	-4	-4	-6
A	-10	-7	-4	-1	0	1	1	-1	-3	-5
A	-12	-9	-6	-3	-2	-1	2	0	-2	-4
T	-14	-11	-8	-5	-4	-1	0	1	-1	-3
G	-16	-13	-10	-7	-6	-3	-2	1	0	0

Global alignment: Needleman-Wunsch

		G	C	C	C	T	A	G	C	G
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
G	-2	1	-1	-3	-5	-7	-9	-11	-13	-15
C	-4	-1	2	0	-2	-4	-6	-8	-10	-12
G	-6	-3	0	1	-1	-3	-5	-5	-7	-9
C	-8	-5	-2	1	2	0	-2	-4	-4	-6
A	-10	-7	-4	-1	0	1	1	-1	-3	-5
A	-12	-9	-6	-3	-2	-1	2	0	-2	-4
T	-14	-11	-8	-5	-4	-1	0	1	-1	-3
G	-16	-13	-10	-7	-6	-3	-2	1	0	0

Best global alignment:

```
GCCCTAGCG
|| | | |
GCGC-AATG
```

alignment core: 0

Local alignment: Smith-Waterman

		G	C	C	C	T	A	G	C	G
	0	0	0	0	0	0	0	0	0	0
G	0	1	0	0	0	0	0	1	0	1
C	0	0	2	1	1	0	0	0	2	0
G	0	1	0	1	0	0	0	1	0	3
C	0	0	2	1	2	0	0	0	2	1
A	0	0	0	1	0	1	1	0	0	1
A	0	0	0	0	0	0	2	0	0	0
T	0	0	0	0	0	1	0	1	0	0
G	0	1	0	0	0	0	0	1	0	1

Local alignment: Smith-Waterman

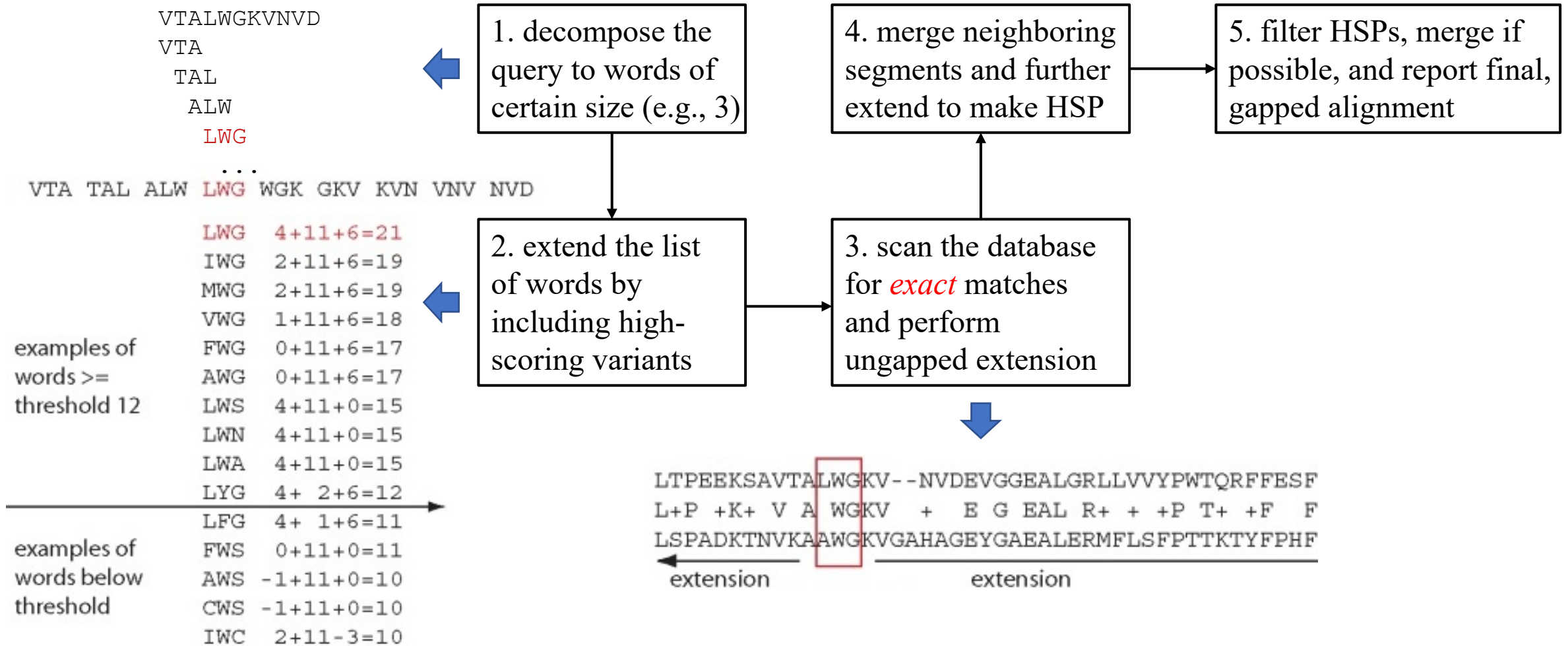|		G	C	C	C	T	A	G	C	G
	0	0	0	0	0	0	0	0	0	0
G	0	1	0	0	0	0	0	1	0	1
C	0	0	2	1	1	0	0	0	2	0
G	0	1	0	1	0	0	0	1	0	3
C	0	0	2	1	2	0	0	0	2	1
A	0	0	0	1	0	1	1	0	0	1
A	0	0	0	0	0	0	2	0	0	0
T	0	0	0	0	0	1	0	1	0	0
G	0	1	0	0	0	0	0	1	0	1

Best local alignment:

GCCCTAGCG
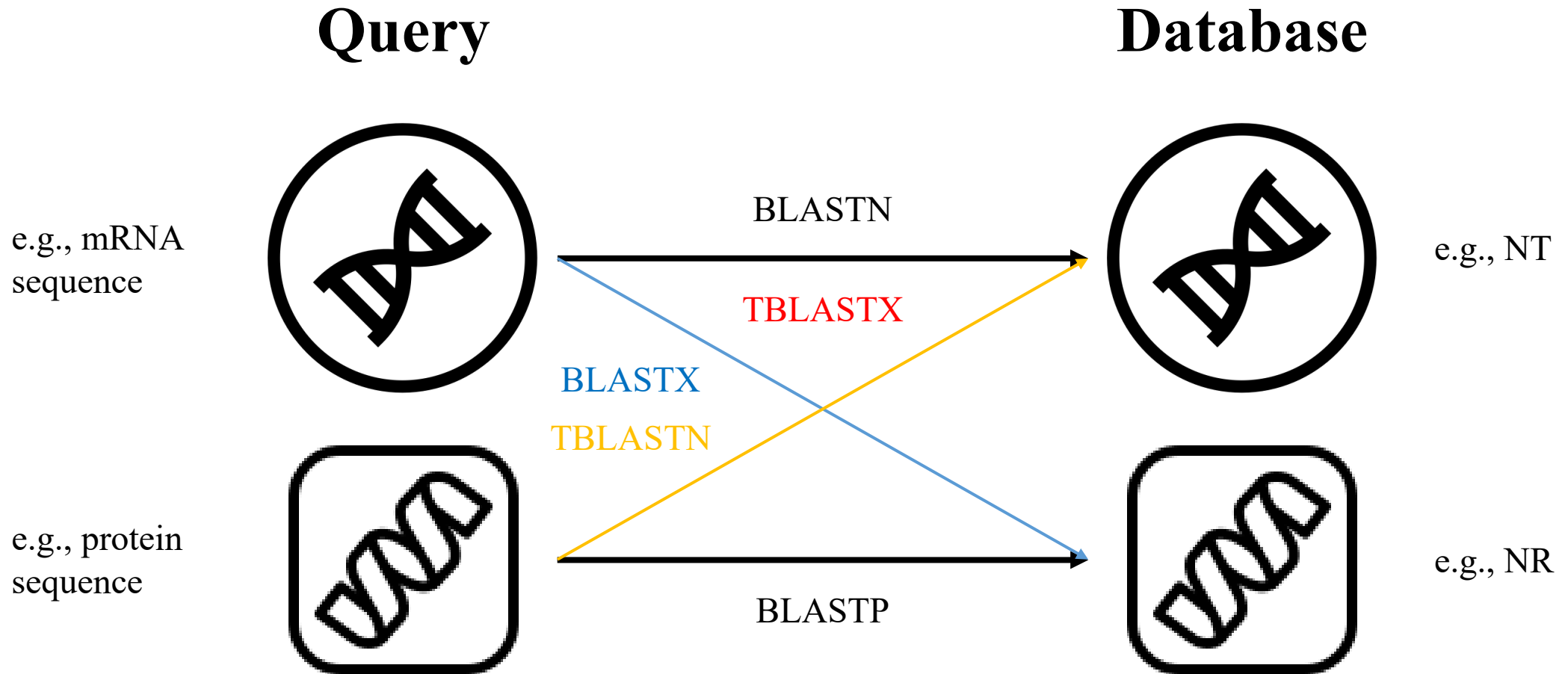 |||
      GCGCAATG

alignment core: 3

BLAST

- **Basic Local Alignment Searching Tool**
  - *"(it) finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance"*.

- One of the most highly cited bioinformatics programs
  - original BLAST (1990):   >83,000 citations
  - gapped BLAST and PSI-BLAST (1997):   >73,000 citations
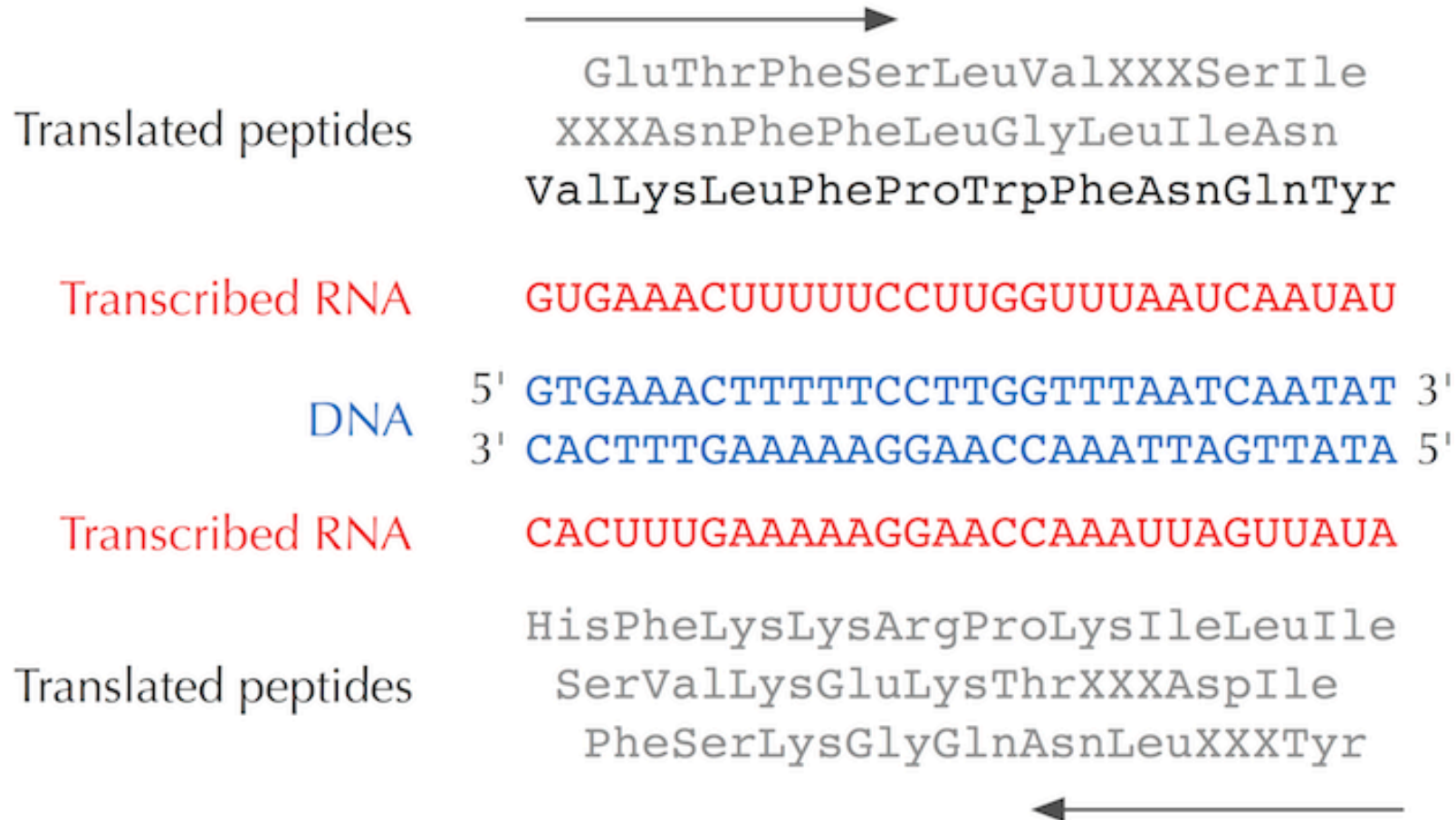 - BLAST+ (2008):   ~7,000 citations

- Why is BLAST so popular?

How does BLAST work?

```
VTALWGKVNVD
VTA
 TAL
  ALW
   LWG
    . . .
VTA  TAL  ALW  LWG  WGK  GKV  KVN  VNV  NVD
```

	LWG	4+11+6=21
	IWG	2+11+6=19
	MWG	2+11+6=19
	VWG	1+11+6=18
examples of	FWG	0+11+6=17
words >=	AWG	0+11+6=17
threshold 12	LWS	4+11+0=15
	LWN	4+11+0=15
	LWA	4+11+0=15
	LYG	4+ 2+6=12
	LFG	4+ 1+6=11
examples of	FWS	0+11+0=11
words below	AWS	-1+11+0=10
threshold	CWS	-1+11+0=10
	IWC	2+11-3=10

1. decompose the query to words of certain size (e.g., 3)

2. extend the list of words by including high-scoring variants

3. scan the database for *exact* matches and perform ungapped extension

4. merge neighboring segments and further extend to make HSP

5. filter HSPs, merge if possible, and report final, gapped alignment

```
LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESF
L+P +K+ V A WGKV   +   E G EAL R+ + +P T+ +F   F
LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF
```

extension ← → extension

Six-frame translation

Translated peptides
```
GluThrPheSerLeuValXXXSerIle
XXXAsnPhePheLeuGlyLeuIleAsn
ValLysLeuPheProTrpPheAsnGlnTyr
```

Transcribed RNA
GUGAAACUUUUUCCUUGGUUUAAUCAAUAU

DNA
5' GTGAAACTTTTTCCTTGGTTTAATCAATAT 3'
3' CACTTTGAAAAAGGAACCAAATTAGTTATA 5'

Transcribed RNA
CACUUUGAAAAAGGAACCAAAUUAGUUAUA

Translated peptides
```
HisPheLysLysArgProLysIleLeuIle
SerValLysGluLysThrXXXAspIle
PheSerLysGlyGlnAsnLeuXXXTyr
```

PSI-BLAST and DELTA-BLAST

- sites in a protein are not all the same

- Positional Specific Scoring Matrix (PSSM)

- achieves higher sensitivity
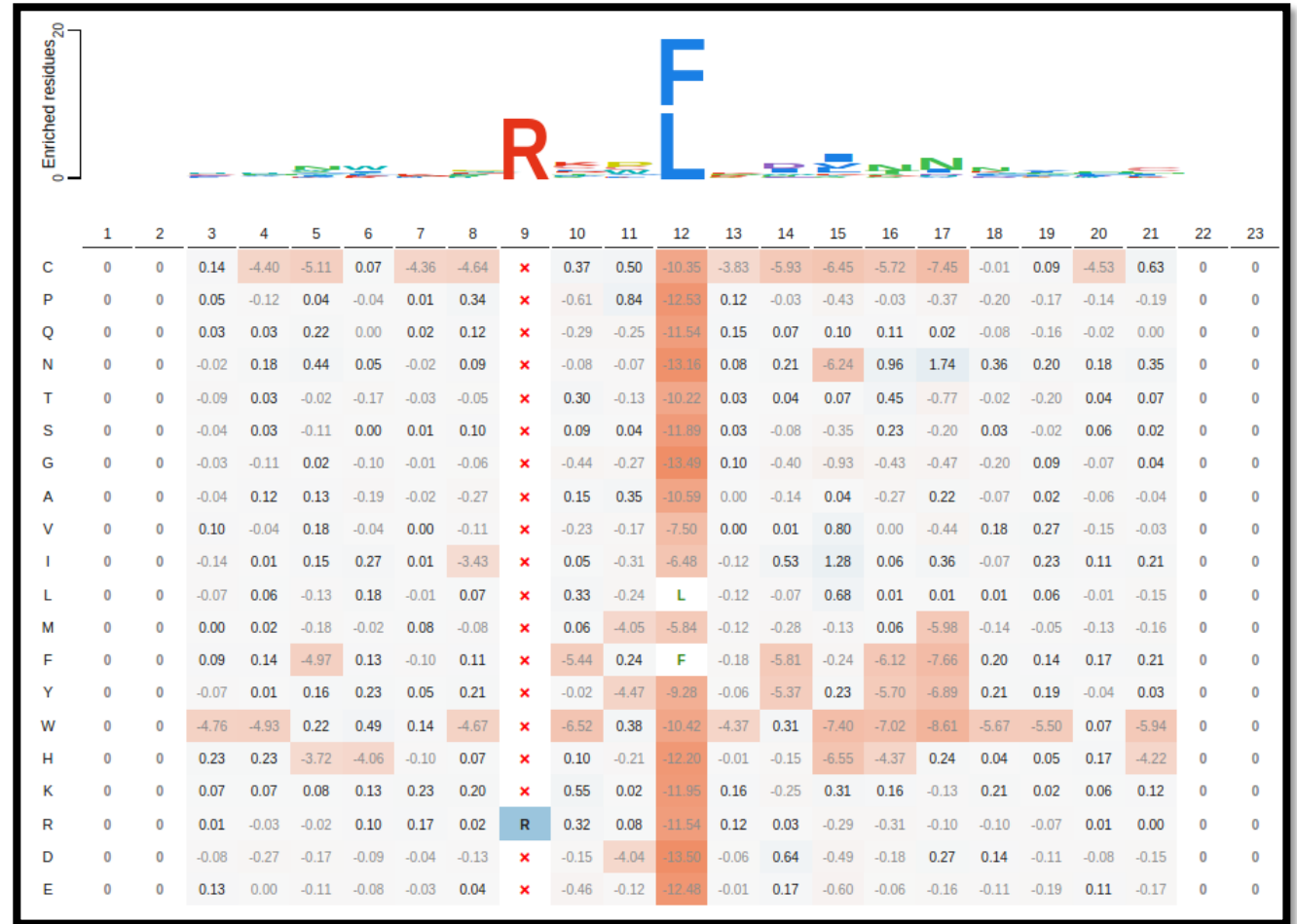
BLAST ®

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.   Learn more

NEWS

A New version of Magic-BLAST(1.5.0) is here.

The BLAST tool for mapping large next-generation RNA or DNA sequencing runs against a whole genome or transcriptome.
Wed, 28 Aug 2019 17:00:00 EST

More BLAST news...

Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein

BLAST Genomes
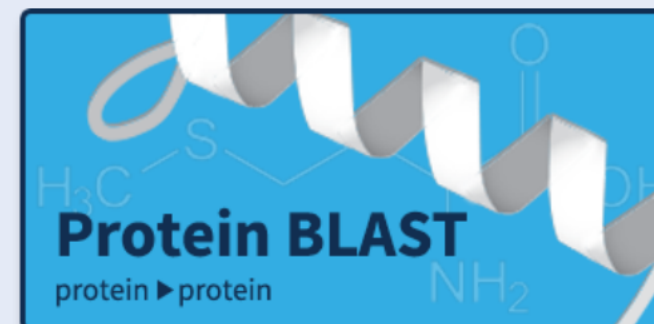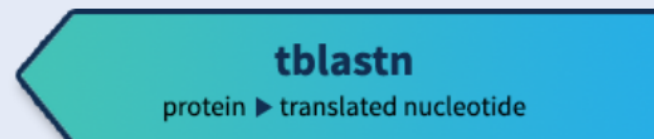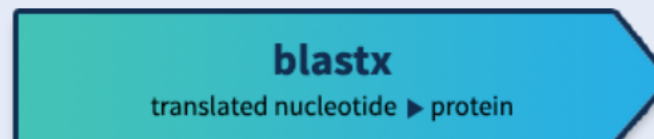
Enter organism common name, scientific name, or tax id   Search

Human   Mouse   Rat   Microbes

○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm ⓘ

BLAST

Search **database nr** using **Blastp (protein-protein BLAST)**

☑ Show results in a new window

⊟ **Algorithm parameters**   Restore default search parameters

General Parameters

Max target sequences	100 ▾	
	Select the maximum number of aligned sequences to display ⓘ	
Short queries	☑ Automatically adjust parameters for short input sequences ⓘ	
Expect threshold	10 ⓘ	
Word size	6 ▾ ⓘ	
Max matches in a query range	0 ⓘ	

Scoring Parameters

Matrix	BLOSUM62 ▾ ⓘ
Gap Costs	Existence: 11 Extension: 1 ▾ ⓘ
Compositional adjustments	Conditional compositional score matrix adjustment ▾ ⓘ

Filters and Masking

Filter	☐ Low complexity regions ⓘ
Mask	☐ Mask for lookup table only ⓘ
	☐ Mask lower case letters ⓘ

BLAST

Search **database nr** using **Blastp (protein-protein BLAST)**

☑ Show results in a new window

Format Request Status

[Formatting options]

Job Title: NP_001243779.1 titin isoform N2BA [Homo sapiens]

Request ID	RGTXK3RT015
Status	Searching
Submitted at	Tue Sep 10 23:19:15 2019
Current time	Tue Sep 10 23:19:26 2019
Time since submission	00:00:10

This page will be automatically updated in **2** seconds

Log in

BLAST ® » blastp suite » results for RID-RGTXK3RT015

Home   Recent Results   Saved Strategies   Help

< Edit Search   Save Search   Search Summary ⌄   ❓ How to read this report?   ▶ BLAST Help Videos   ↻ Back to Traditional Results Page

ℹ️   Your search is limited to records that include: Mus (taxid:10088)

Job Title	**NP_001243779.1 titin isoform N2BA [Homo sapiens]**
RID	RGTXK3RT015 *Search expires on 09-12 11:19 am* Download All ⌄
Program	BLASTP ❓ Citation ⌄
Database	nr See details ⌄
Query ID	lcl|Query_260495
Description	NP_001243779.1 titin isoform N2BA [Homo sapiens]
Molecule type	amino acid
Query Length	34350
Other reports	Distance tree of results Multiple alignment MSA viewer ❓

Filter Results

Organism *only top 20 will appear*   ☐ exclude

```
Type common  name, binomial, taxid or group  name
```

➕ Add organism

Percent Identity   **E value**

[   ] to [   ]   [   ] to [   ]

Filter   Reset

Descriptions   **Graphic Summary**   **Alignments**   **Taxonomy**

Sequences producing significant alignments

Download ⌄   Manage Columns ⌄   Show [100 ⌄]   ❓

☑ select all   100 sequences selected   GenPept   Graphics   Distance tree of results   Multiple alignment

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	titin isoform N2-A [Mus musculus]	44050	1.270e+05	96%	0.0	93.16%	NP_035782.3
☑	titin isoform X47 [Mus caroli]	43983	67215	93%	0.0	93.03%	XP_029329032.1
☑	titin isoform X3 [Mus pahari]	43769	1.126e+05	100%	0.0	93.90%	XP_029391868.1
☑	titin isoform X1 [Mus musculus]	43745	1.260e+05	100%	0.0	92.66%	XP_017172628.1
☑	RecName: Full=Titin; AltName: Full=Connectin [Mus musculus]	43735	1.163e+05	100%	0.0	92.86%	A2ASS6.1
☑	titin isoform X37 [Mus caroli]	43699	89404	98%	0.0	92.43%	XP_029328975.1
☑	titin isoform X42 [Mus caroli]	43697	1.093e+05	98%	0.0	92.43%	XP_029329006.1
☑	titin isoform X39 [Mus caroli]	43697	1.116e+05	98%	0.0	92.43%	XP_029328983.1
☑	titin isoform X40 [Mus caroli]	43697	1.113e+05	98%	0.0	92.43%	XP_029328994.1
☑	titin [Mus musculus]	43693	80634	99%	0.0	93.54%	EDL27217.1
☑	titin isoform X4 [Mus caroli]	43690	89120	98%	0.0	92.55%	XP_029328892.1
☑	titin isoform X35 [Mus caroli]	43690	1.158e+05	98%	0.0	92.43%	XP_029328964.1

▣ Feedback

BLAST ® » blastp suite » results for RID-RGTXK3RT015

< **Edit Search**   Save Search   Search Summary ⌄

❓ How to read this report?   ▶ BLAST Help Videos   ↩ Back to Traditional Results Page

ℹ Your search is limited to records that include: Mus (taxid:10088)

Job Title	NP_001243779.1 titin isoform N2BA [Homo sapiens]
RID	RGTXK3RT015 Search expires on 09-12 11:19 am Download All ⌄
Program	BLASTP ❓ Citation ⌄
Database	nr See details ⌄
Query ID	lcl|Query_260495
Description	NP_001243779.1 titin isoform N2BA [Homo sapiens]
Molecule type	amino acid
Query Length	34350
Other reports	Distance tree of results Multiple alignment MSA viewer ❓

Filter Results

Organism *only top 20 will appear*   ☐ exclude

[Type common name, binomial, taxid or group name]

➕ Add organism

Percent Identity   **E value**

[   ] to [   ]   [   ] to [   ]

Filter   **Reset**

Descriptions	**Graphic Summary**	Alignments	Taxonomy

👆 hover to see the title   ➤ click to show alignments   ☑ Show Conserved Domains

Alignment Scores   ■ < 40   ■ 40 - 50   ■ 50 - 80   ■ 80 - 200   ■ >= 200   ❓
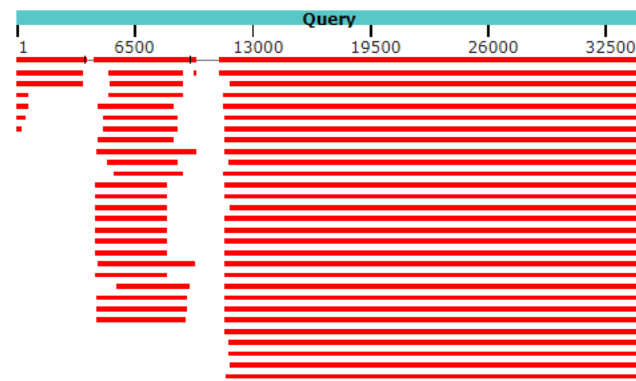
100 sequences selected ❓

Hit the button to
See conserved domains from CDD

Distribution of the top 200 Blast Hits on 100 subject sequences

BLAST ® » blastp suite » results for RID-RGTXK3RT015

< Edit Search   Save Search   Search Summary ⌄   ❓ How to read this report?   ▶ BLAST Help Videos   ↺ Back to Traditional Results Page

ℹ Your search is limited to records that include: Mus (taxid:10088)

Job Title	NP_001243779.1 titin isoform N2BA [Homo sapiens]
RID	RGTXK3RT015 Search expires on 09-12 11:19 am Download All ⌄
Program	BLASTP ❓ Citation ⌄
Database	nr See details ⌄
Query ID	lcl|Query_260495
Description	NP_001243779.1 titin isoform N2BA [Homo sapiens]
Molecule type	amino acid
Query Length	34350
Other reports	Distance tree of results Multiple alignment MSA viewer ❓

Filter Results

Organism only top 20 will appear   ☐ exclude

[Type common name, binomial, taxid or group name]

➕ Add organism

Percent Identity   **E value**

[   ] to [   ]   [   ] to [   ]

Filter   Reset

Descriptions   Graphic Summary   **Alignments**   Taxonomy

Alignment view [Pairwise   ⌄] ❓   Download ⌄

100 sequences selected ❓

⬇ Download ⌄   GenPept Graphics   Sort by [E value   ⌄]   ▼ Next ▲ Previous ◀◀Descriptions

titin isoform N2-A [Mus musculus]

Sequence ID: NP_035782.3   Length: 33467   Number of Matches: 813

Range 1: 10251 to 33467 GenPept Graphics   ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
44050 bits(114345)	0.0	Compositional matrix adjust.	21681/23273(93%)	22507/23273(96%)	112/23273(0%)

```
Query  11134  APPAKVPEVPKKVEEKRIILPKEEEVLPVEV-------TEEPEEEPISEEEIPEEPPSIE  11186
              APP KVPEV KKVEE+RII PKEEEV P EV       T E  E    E  E
Sbjct  10251  APPTKVPEVSKKVEERRIIPKEEEVPPAEVYEEAEEPTPEEIPEEPPSIEEEEIVEEEE  10310

Query  11187  EVEEVAPPRVPEVIKKAVPEAPTPVPKKVEAPPAKVSKKIPEEKVPVPVQKKEAPPAKVP  11246
              E EEV PPR PEV+KKAVPEAPTPVPKK EAPPAKV KKIPEEKVPVPVQKKEAPPAKVP
Sbjct  10311  EEEEVLPPRAPEVVKKAVPEAPTPVPKKAEAPPAKVPKKIPEEKVPVPVQKKEAPPAKVP  10370

Query  11247  EVPKKVPEKK------VLVPKKEAVPPAKGRTVLEEKVSVAFRQEVVVKERLELEVVEAE  11300
              EVPKKVPEKK      V VPKKEAVPPAKG+ V EEK+SVA++QE +V+ER+ELE VEA+
Sbjct  10371  EVPKKVPEKKKIPEKKVPVPKKEAVPPAKGKAVFEEKISVAYQQEELVQERIELELVEAK  10430
```

Related Information

Gene - associated gene details

Genome Data Viewer - aligned genomic context

BLAST ® » blastp suite » results for RID-RGTXK3RT015

‹ Edit Search   Save Search   Search Summary ⌄   How to read this report?   ▶ BLAST Help Videos   ↺ Back to Traditional Results Page

ℹ Your search is limited to records that include: Mus (taxid:10088)

Job Title	**NP_001243779.1 titin isoform N2BA [Homo sapiens]**
RID	RGTXK3RT015 *Search expires on 09-12 11:19 am* Download All ⌄
Program	BLASTP ❓ Citation ⌄
Database	nr See details ⌄
Query ID	lcl|Query_260495
Description	NP_001243779.1 titin isoform N2BA [Homo sapiens]
Molecule type	amino acid
Query Length	34350
Other reports	Distance tree of results Multiple alignment MSA viewer ❓

Filter Results

Organism   *only top 20 will appear*   ☐ exclude

```
Type common name, binomial, taxid or group name
```

➕ Add organism

Percent Identity   **E value**

[   ] to [   ]   [   ] to [   ]

Filter   Reset

Descriptions	Graphic Summary	Alignments	**Taxonomy**

Reports | **Lineage** | **Organism** | **Taxonomy** |

100 sequences selected ❓

Organism	Blast Name	Score	Number of Hits	Description
Mus	rodents		101	
. Mus	rodents		93	
. . Mus musculus	rodents	44050	34	Mus musculus hits
. . Mus caroli	rodents	43983	59	Mus caroli hits
. Mus pahari	rodents	43769	8	Mus pahari hits

Blog   ❓ Support Center

Feedback

Critical BLAST parameters

- Scoring scheme:
  - scoring matrix
  - gap costs

- Word size

- *E*-value

How to interpret BLAST results

- Expect value: "*the number of high-scoring segment pairs (HSPs) expected to occur with a score of at least S.*"

Karlin Altschul statistics

$$E = Kmn\mathrm{e}^{-\lambda S} \longrightarrow \text{alignment score}$$

$$\boxed{E = -\ln(1-P)}$$

lengths of the query and the database

Questions?